PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 5 : C12Q 1/68

A1

(11) International Publication Number: WO 90/04652

(43) International Publication Date: 3 May 1990 (03.05.90)

(21) International Application Number: PCT/US89/04741

(22) International Filing Date: 23 October 1989 (23.10.89)

(30) Priority data:
261,702 24 October 1988 (24.10.88) US

(71) Applicant: DNAX RESEARCH INSTITUTE OF MOLECULAR AND CELLULAR BIOLOGY, INC. [US/US]; 901 California Avenue, Palo Alto, CA 94304-1104 (US).

(72) Inventor: MACEVICZ, Stephen, C. ; 1034 Innsbruck Street, Livermore, CA 94550 (US).

(74) Common Representatives: MACEVICZ, Stephen, C. et al.; DNAX Research Institute of Molecular and Cellular Biology, Inc., 901 California Avenue, Palo Alto, CA 94304-1104 (US).

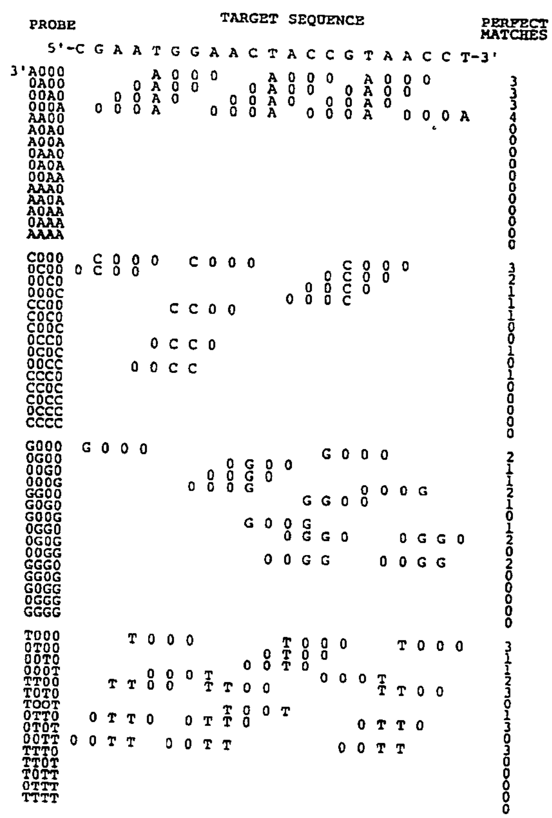
(81) Designated States: AT (European patent), BE (European patent), CH (European patent), DE (European patent), FR (European patent), GB (European patent), IT (European patent), JP, LU (European patent), NL (European patent), SE (European patent).

Published
With international search report.
Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(54) Title: DNA SEQUENCING BY MULTIPLE MIXED OLIGONUCLEOTIDE PROBES

(57) Abstract

A method is provided for sequencing nucleic acids without the need to separate similarly sized DNAs or RNAs by gel electrophoresis. The method relies on the separate hybridization of multiple mixed oligonucleotide probes to a target sequence. The mixed oligonucleotide probes comprise sequences of fixed and non-fixed bases corresponding to every possible permutation of fixed and non-fixed bases less than or equal to the length of the probes. For each probe, the hybridizations provide the number of times the probe's particular sequence of fixed bases appears in the target sequence. The target sequence is then mathematically reconstructed from this data and a knowledge of the probe sequences.



CGK00033901

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AT	Austria	ES	Spain	MG	Madagascar
AU	Australia	FI	Finland	ML	Mali
BB	Barbados	FR	France	MR	Mauritania
BE	Belgium	GA	Gabon	MW	Malawi
BF	Burkina Fasso	GB	United Kingdom	NL	Netherlands
BG	Bulgaria	HU	Hungary	NO	Norway
BJ	Benin	IT	Italy	RO	Romania
BR	Brazil	JP	Japan	SD	Sudan
CA	Canada	KP	Democratic People's Republic of Korea	SE	Sweden
CF	Central African Republic	KR	Republic of Korea	SN	Senegal
CG	Congo	LI	Liechtenstein	SU	Soviet Union
CH	Switzerland	LK	Sri Lanka	TD	Chad
CM	Cameroon	LU	Luxembourg	TG	Togo
DE	Germany, Federal Republic of	MC	Monaco	US	United States of America
DK	Denmark				

DNA SEQUENCING BY MULTIPLE MIXED OLIGONUCLEOTIDE PROBES

Background

The ability to determine DNA sequences is crucial for understanding the function and control of genes and for applying many of the basic techniques of molecular biology. Native DNA consists of two linear polymers, or strands of nucleotides. Each strand is a chain of nucleosides linked by phosphodiester bonds. The two strands are held together in an antiparallel orientation by hydrogen bonds between complementary bases of the nucleotides of the two strands: deoxyadenosine (A) pairs with thymidine (T) and deoxyguanosine (G) pairs with deoxycytidine (C).

Presently there are two basic approaches to DNA sequence determination: the dideoxy chain termination method, e.g. Sanger et al, Proc. Natl. Acad. Sci., Vol. 74, pgs. 5463-5467 (1977); and the chemical degradation method, e.g. Maxam et al, Proc. Natl. Acad. Sci., Vol. 74, pgs.560-564 (1977). The chain termination method has been improved in several ways, and serves as the basis for all currently available automated DNA sequencing machines, e.g. Sanger et al, J. Mol. Biol., Vol. 143, pgs. 161-178 (1980); Smith et al, Nucleic Acids Research, Vol. 13, pgs. 2399-2412 (1985); Smith et al, Nature, Vol. 321, pgs. 674-679 (1987); Prober et al, Science, Vol. 238, pgs. 336-341 (1987), Section II, Meth. Enzymol., Vol. 155, pgs. 51-334 (1987),and Church et al, Science, Vol 240, pgs. 185-188 (1988).

Both the chain termination and chemical degradation methods require the generation of one or more sets of labeled DNA fragments, each having a common origin and each terminating with a known base. The set or sets of fragments must then be separated by size to obtain sequence information. In both methods, the DNA fragments are separated by high resolution gel electrophoresis. Unfortunately, this step severely limits the size of the DNA chain that can be sequenced at one time. Non-automated sequencing can

accommodate a DNA chain of up to about 500 bases under optimal conditions, and automated sequencing can accommodate a chain of up to about 300 bases under optimal conditions, Bankier et al, Meth. Enzymol., Vol. 155, pgs. 51-93 (1987); Roberts, Science, Vol. 238, pgs. 271-273 (1987); and Smith et al, Biotechnology, Vol. 5, pgs. 933-939 (1987).

This limitation represents a major bottleneck for many important medical, scientific, and industrial projects aimed at unraveling the molecular structure of large regions of plant or animal genomes, such as the project to sequence all or major portions of the human genome, Smith et al, Biotechnology (cited above).

In addition to DNA sequencing, nucleic acid hybridization has also been a crucial element of many techniques in molecular biology, e.g. Hames et al, eds., Nucleic Acid Hybridization: A Practical Approach (IRL Press, Washington, D.C., 1985). In particular, hybridization techniques have been used to select rare cDNA or genomic clones from large libraries by way of mixed oligonucleotide probes, e.g. Wallace et al, Nucleic Acids Research, Vol. 6, pgs. 3543-3557 (1979), or by way of interspecies probes, e.g. Gray et al, Proc. Natl. Acad. Sci., Vol. 80, pgs. 5842-5846 (1983). Nucleic acid hybridization has also been used to determine the degree of homology between sequences, e.g. Kafatos et al, Nucleic Acids Research, Vol. 7, pgs. 1541-1552 (1979), and to detect consensus sequences, e.g. Oliphant et al, Meth. Enzymol., Vol. 155, pgs. 568-582 (1987). Implicit to all of these applications is the notion that the known probe sequences contain information about the unknown target sequences. This notion apparently has never been exploited to obtain detailed sequence information about a target nucleic acid. In view of the limitations of current DNA sequencing methods, it would be advantageous for the scientific and industrial communities to have available an alternative method for sequencing DNA which (1) did not require gel electrophoretic separation of similarly sized DNA fragments, (2) had the capability of providing the sequence of very long

DNA chains in a single operation, and (3) was amenable to automation.

Disclosure of the Invention

The invention is directed to a method for determining the nucleotide sequence of a DNA or an RNA molecule using multiple mixed oligonucleotide probes. Sequence information is obtained by carrying out a series of hybridizations whose results provide for each probe the number of times the complement of the probe's sequence occurs in the RNA or DNA whose sequence is to be determined. The nucleotide sequence of the RNA or DNA is reconstructed from this information and from a knowledge of the probes' sequences. The nucleic acid whose sequence is to be determined is referred to herein as the target sequence.

The mixed oligonucleotide probes of the invention are selected from a set whose members' sequences include every possible complementary sequence to subsequences of a predetermined length within the target sequence. The series of hybridizations are separately carried out such that one or more of the probes selected from the set are combined with known quantities of the target sequence, e.g. on a nitrocellulose filter, or like substrate, under conditions which substantially allow only perfectly matched probe sequences to hybridize with the target sequence. Probe sequences having mismatched bases are substantially removed, e.g. by washing, and the quantity of perfectly matched probe remaining hybridized to the target sequence is determined.

In one embodiment of the invention, the set of probes comprises four subsets. Each of the four subsets contains probes representing every possible sequence, with respect to the size of the probe (which is predetermined), of only one of the four bases. For example, the first subset can contain probes where every possible sequence of G is represented; the second subset can contain probes where every possible sequence of T is represented; and so on for C and A. If the probes

were each 8 bases long, a member probe of the adenosine subset can be represented as follows:

$$3'\text{-}AA\begin{pmatrix}C\\G\\T\end{pmatrix}\begin{pmatrix}C\\G\\T\end{pmatrix}A\begin{pmatrix}C\\G\\T\end{pmatrix}\begin{pmatrix}C\\G\\T\end{pmatrix}A\text{-}5'$$

Formula I

The symbol $\begin{pmatrix}C\\G\\T\end{pmatrix}$ means that any of the bases C, G, or T may occupy the position where the symbol is located. Thus, the above probe has a multiplicity, or degeneracy, of 1x1x3x3x1x3x3x1, or 81. When it is clear from the context which subset is being considered, the above notation will be simplified to AA00A00A, where A represents deoxyadenosine and 0 represents the absence of deoxyadenosine.

Preferably, base analogs are employed in the oligonucleotide probes whose base pairing characteristics permit one to reduce the multiplicity of the probe. For example, in the probe of Formula II, because deoxyinosine (I) forms nearly equally strong base pairs with A and C, but forms only a weak or destabilizing base pair with either G or T, deoxyinosine can replace G and T in the probe, Martin et al, Nucleic Acids Research, Vol. 13, pgs. 8927-8938 (1985). Thus, a probe equivalent to that of Formula I, but which has a much lower multiplicity (i.e. only 16) can be represented as follows:

$$3'\text{-}AA\begin{pmatrix}I\\C\end{pmatrix}\begin{pmatrix}I\\C\end{pmatrix}A\begin{pmatrix}I\\C\end{pmatrix}\begin{pmatrix}I\\C\end{pmatrix}A\text{-}5'$$

Formula II

Generally, base analogs are preferred which form strong base pairs (i.e., comparable in binding energy to the natural base pairs) with two or three of the four natural bases, and a weak or destabilizing base pair with the complement of a fixed base (defined below). Such base analogs are referred to

herein as degeneracy-reducing analogs.

It is not critical that the probes all have the same length, although it is important that they have known lengths and that their sequences be predetermined. Generally, the probes will be fixed at a predetermined number of positions with known bases (not necessarily of the same kind), e.g. as the A in Formula I, and the remaining positions will each be filled by a base randomly selected from a predetermined set, e.g. T, G, and C as in Formula I, or I and C as in Formula II. The positions in a probe which are non-degenerate in their base pairing, i.e. have only a single natural base, are referred to herein as fixed positions. The bases occupying fixed positions are referred to herein as fixed bases. For example, the fixed bases in the probes of Formulas I and II are deoxyadenosine at positions one, two, five, and eight with respect to the 3' end of the probe.

Generally, sets and/or subsets of the invention each contain at least one probe having a sequence of fixed and non-fixed positions equivalent to that of each permutation of a plurality of fixed and non-fixed positions less than or equal to the length of the probe. That is, an important feature of the invention is that the probes collectively contain subsequences (up to the total length of the probe) which correspond to every possible permutation of fixed and non-fixed positions of each of a plurality of combinations of fixed and non-fixed positions, the plurality including combinations containing from zero to all fixed positions. For example, consider a subset of probes of the invention that consists of 8-mer probes whose fixed positions contain only deoxyadenosine <u>and</u> whose initial (i.e., 3'-most) position is fixed. The probes of Formulas I and II are members of such a subset. Within such a subset, there is a least one probe having a subsequence of fixed and non-fixed positions in positions 2 through 8 which corresponds to each possible permutation of fixed and non-fixed positions for subsequences having no fixed positions (one such permutation: A0000000), one fixed position (seven such permutations, e.g. A000A000),

two fixed positions (twenty-one such permutations, e.g. A00AA000), three fixed positions (thirty-five such permutations, e.g. A0000AAA), four fixed positions (thirty-five such permutations, e.g. A0AAAA00), five fixed positions (twenty-one such permutations, e.g. AAA00AAA), six fixed positions (seven such permutations, e.g. AAAA0AAA), and seven fixed positions (one such permutation: AAAAAAAA). Thus, the subset has at least 1 + 7 + 21 + 35 + 35 + 21 + 7 + 1 = 128 members.

The presence of one or more predetermined known sequence regions in the target sequence facilitates the reconstruction of the target sequence. Accordingly, in a preferred embodiment, the target sequence contains one or more regions of known sequence, these regions being referred to herein as known sequence regions. More preferably, the target sequence contains a first and a second known sequence region, the first and second known sequence regions being positioned on opposite ends of the region of the target sequence containing the unknown sequence of nucleotides. This unknown sequence of nucleotides is referred to herein as the unknown sequence region. Most preferably, the first and second known sequence regions are at least the length of the longest probe sequence.

I. Composition and Labeling of the Probes

Mixed oligonucleotide probes for the invention are preferably synthesized using an automated DNA synthesizer, e.g. Applied Biosystems (Foster City, CA) models 381A or 380B, or like instrument. At non-fixed positions mixtures of the appropriate nucleotide precursors are reacted with the growing oligonucleotide chain so that oligonucleotides having different different bases at that position are synthesized simultaneously, e.g. as disclosed by Wallace et al, Nucleic Acids Research, Vol. 6, pgs.3543-3557 (1979), and Oliphant et al, Meth. Enzymol., Vol. 155, pgs. 568-582 (1987). The probes may be synthesized by way of any of the available chemistries, e.g. phosphite triester, Beaucage et al,

Tetrahedron Letters, Vol. 22, pgs. 1859-1862 (1981); Caruthers et al, U.S. patents 4,415,723, 4,458,066, and 4,500,707; phosphotriester, Itakura, U.S. patent 4,401,796; hydrogen phosphonate, e.g. Froehler et al, *Nucleic Acids Research*, Vol. 14, pgs. 5399-5407 (1986); or the like. Once synthesized, the oligonucleotides are purified for labeling by well known techniques, usually HPLC or gel electrophoresis, e.g. *Applied Biosystems DNA Synthesizer Users Bulletin*, Issue No. 13-Revised (April 1, 1987).

Selecting the lengths of the probes is an important aspect of the invention. Several factors influence the choice of length for a given application, including (1) the ease with which hybridization conditions can be manipulated for preferentially hybridizing probes perfectly matched to the target sequence, (2) the ability to distinguish between roughly integral amounts of perfectly matched probe hybidized to the target sequence (e.g. if the probe is relatively long so that the expected frequency of probe sequences perfectly complementary to the target is low, one may be required to distinguish (for example) between amounts of probe in the range of 10, 20, or 30 picomoles--to infer that 1, 2, or 3 copies of the probe are present on the target; if the probe is relatively short so that the expected frequency of probe sequences perfectly complementary to the target is high, one may be required to distinguish (for example) between amounts of probe in the range of 110, 120, or 130 picomoles--to infer that 11, 12, or 13 copies of the probe are present on the target; since the fractional differences between the latter quantities are small, there may be less confidence in the inferred copy number); (3) whether probe multiplicity permits hybridization with reasonable Cot values (longer probes are more degenerate than shorter probes, and require higher Cot values for hybridization, the converse is true of shorter probes); (4) the practicality of carrying out separate hybridizations for each type of probe (longer probes give rise to larger sets of probes, as described above); and (5) the tractability of the sequence reconstruction problem (the greater the number of

copies of each probe type on the target sequence--which is the tendency if shorter probes are employed, the more difficult the reconstruction problem). Probe sizes in the range of 7 to 11 bases are preferred. More preferably, probes sizes are in the range of 8 to 10 bases, and most preferably, probe sizes are in the range of 8 to 9 bases.

Preferably, degeneracy-reducing analogs are employed at the non-fixed positions of the probes to reduce probe multiplicity, or degeneracy. Many synthetic and natural nucleoside and nucleotide analogs are available for this purpose, e.g. Scheit, Nucleotide Analogs (John Wiley & Sons, New York, 1980). For example, degeneracy-reducing analogs include deoxyinosine for use in cytosine or adenosine probes to replace G and T at non-fixed positions, 2-aminopurine for use in cytosine or guanosine probes to replace A and T at non-fixed positions, and N^4- methoxydeoxycytidine, N^4-aminodeoxycytidine, or 5-fluorodeoxyuridine for use in adenosine or guanosine probes to replace T and C. Use of deoxyinosine in oligonucleotide probes is disclosed by Martin et al (cited above); Seela et al, Nucleic Acids Research, Vol. 14, pgs. 1825-1844 (1986); Kawase et al, Nucleic Acids Research, Vol. 14, pgs. 7727-7737 (1986); Ohtsuka et al, J. Biol. Chem., Vol. 260, pgs. 2605-2608 (1985); and Takahashi et al, Proc. Natl. Acad. Sci., Vol. 82. pgs. 1931-1935 (1985). Deoxyinosine phosphoramidite precursors for automated DNA synthesis are available commercially, e.g. Applied Biosystems (Foster City, CA). The synthesis of N^4-methoxycytidine and its incorporation into oligonucleotide probes is disclosed by Anand et al, Nucleic Acids Research, Vol. 15, pgs. 8167-8176 (1987). The synthesis of 2-aminopurine and its incorporation into oligonucleotide probes is disclosed by Eritja et al, Nucleic Acids Research, Vol. 14, pgs. 5869-5884 (1986). The synthesis of 5-fluorodeoxyuridine and its incorporation into oligonucleotide probes is disclosed by Habener et al, Proc. Natl. Acad. Sci., Vol. 85, pgs. 1735-1739 (1988). And the preparation of N^4-aminodeoxycytidine is disclosed by

Negishi et al, Nucleic Acids Research, Vol. 11, pgs. 5223-5233 (1983).

Nucleoside analogs are also employed in the invention to reduce the differences in binding energies between the various complementary bases. In particular, 2-aminoadenine can replace thymine in either the probe or target sequences to reduce the binding energy differences between A-T nucleoside pairs and G-C nucleoside pairs, e.g. Kirnos et al (cited above), Chollet et al (cited above), and Cheong et al, Nucleic Acids Research, Vol. 16, pgs. 5115-5122 (1988). Procedures for synthesizing oligonucleotides containing 2-aminoadenine are disclosed by Chollet et al (cited above); Gaffney et al, Tetrahedron, Vol. 40, pgs. 3-13 (1984), and Chollet et al Chemica Scripta, Vol. 26, pgs. 37-40 (1986). Likewise, 2-amino-2'-deoxyadenosine can replace deoxyadenosine to increase the binding energy at positions where A-T pairs occur, e.g. Huynh-Dihn et al, Proc. Natl. Acad. Sci., Vol. 82, pgs. 7510-7514 (1985).

In some embodiments, it may be preferable to replace a more degenerate probe with several less degenerate probes which collectively are capable of obtaining the same information about the target sequence. For example, consider the 9-mer probe A00000000. This probe can be replaced by the three less degenerate probes A0000000C, A0000000G, and A0000000T. Thus, at the cost of two additional hybridizations, the degeneracy of the most degenerate probe in the set is reduced from 256 to 128 (assuming the use of deoxyinosine at non-fixed positions).

The oligonucleotides of the invention can be labeled in a variety of ways to form probes, including the direct or indirect attachment of radioactive moieties, fluorescent moieties, electron dense moieties, and the like. It is only important that each sequence within a probe be capable of generating a signal of the same magnitude, so that quantitative measurements of probe number can be made. There are several means available for derivatizing oligonucleotides with reactive functionalities which, permit the addition of a label, e.g. Connolly, Nulceic Acids

Research, Vol 15, pgs. 3131-3139 (1987); Gibson et al, Nucleic Acids Research, Vol. 15, pgs. 6455-6467 (1987); Spoat et al, Nucleic Acids Research, Vol. 15, pgs. 4837-4848 (1987); and Mathews et al, Anal. Biochem., Vol. 169, pgs. 1-25 (1988).

In one preferred embodiment, the oligonucleotides of the invention are radioactively labeled with ^{32}P using standard protocols, e.g. Maxim and Gilbert, Meth. Enzymol., Vol. 65, pgs. 499-560 (1980). ^{32}P-labeled probes of the invention are preferably applied to targent DNAs anchored to nitrocellulose, nylon, or the like, at a concentration in the range of about 1-10 ng/ml, and more preferably, in the range of about 1-5 ng/ml. The specific activities of the probe are preferably in the range of about 1-5 X 10^6 cpm/ml.

II. Hybridization

The hybridizations of the probes to the target sequence are carried out in a manner which allows mismatched probe sequences and nonspecifically bound probe sequences to be separated from the duplexes formed between the perfectly matched probe sequences and the target sequence. Usually the separation is carried out by a washing step. Preferably, the first step in the hybridizations is to anchor the target sequence so that washes and other treatments can take place with minimal loss of the target sequences. The method selected for anchoring the target sequence depends on several factors, including the length of the target, the method used to prepare copies of the target, and the like. Preferably, the target sequence is anchored by attaching it to a substrate or solid phase support, such as nitrocellulose, nylon-66, or the like, or such as derivatized microspheres, e.g. Kremsky et al, Nucleic Acids Research, Vol. 15, pgs.2891-2909 (1987).

A known quantity of single or double stranded copies of the target sequence is anchored to the substrate, or solid phase support. As used herein "known quantity" means amounts from which integral numbers of perfectly matched probes can be determined. In some embodiments this means known gram or molar quantities of the target sequence. In

other embodiments, it can mean equal amounts of target sequence on the plurality of solid phase supports, so that signals corresponding to integral numbers of probes can be discerned by comparing signals from the plurality of supports, or by comparing signals to specially provided standards. Preferably, the anchoring means is loaded to capacity with the target sequence so that maximal signals are produced after hybridization. The target sequence can be prepared in double stranded form, denatured, and then applied to the anchoring means, which is preferably a solid phase support, such as nitrocellulose, GeneScreen, or the like. When the target sequence is prepared in double stranded form, it is preferably excised from its cloning vector with one or more endonucleases which leave blunt ended fragments, e.g. Eco RV, Alu I, Bal I, Dra I, Nae I, Sma I, or the like. In this case, both the coding, or sense, strand and the noncoding, or antisense, strand are sequenced simultaneously. Because of sequence complementarity, the reconstruction problem is no more difficult than in the single stranded case.

Suitable vectors for preparing double stranded target sequences are those of the pUC series, e.g. Yanisch-Perron et al, Gene, Vol. 33, pgs. 103-119 (1985). These vectors are readily modified by adding unique restriction sites to their polylinker regions. The new unique restriction sites are selected from restriction endonucleases that leave flush-ended fragments after digestion. For example, chemically synthesized fragments containing such sites can be inserted into the Hind III and Eco RI sites of pUC18 or pUC19. For these vectors such sites include Bal I, Eco RV, Hpa I, Nae I, Nru I, Stu I, Sna BI, and Xca I. With the modified pUC, the precursor of the target sequence (i.e. the unknown sequence region) can be inserted into a preexisting polylinker site, e.g. Bam HI; the vector can be amplified and isolated; and the target sequence can be excised via the restriction endonucleases that leave flush-ended fragments. The fragments of the polylinker region excised along with the unknown sequence region then become the known sequence regions of the

target sequence.

The preferred method of anchoring DNA to nitrocellulose filters is essentially that described by Kafatos et al (cited above). Up to about 1 ug of target sequence is applied per square millimeter of the filter. Before application the DNA is denatured, preferably in 0.3 to 0.4 N NaOH for about 10 minutes, after which it is chilled with an equal volume of cold water, or optionally cold 2 M ammonium acetate, to a concentration of about 16 ug/ml. Known quantities of the denatured target are spotted onto the filter by carefully controlling the volume of liquid deposited. After each sample is spotted (approximately 1.5 minutes), the filter can optionally be rinsed through with a drop of 1 M ammonium acetate containing about .02-0.2 N NaOH, pH 7.8-9.0. Filters may also be washed with 4xSSC (defined below), e.g. about 200 ml. The filters are air dried, shaken in 2x Denhardt's solution (defined below) for at least 1 hour, drained and air dried again, and baked under vacuum at 80° C for about 2 hours.

Hybridization of the probes to the target sequence usually comprises three steps: a prehybridization treatment, application of the probe, and washing. The purpose of the prehybridization treatment is to reduce nonspecific binding of the probe to the anchoring means and non-target nucleic acids. This is usually accomplished by blocking potential nonspecific binding sites with blocking agents such as proteins, e.g. serum albumin (a major ingredient of Denhardt's solution). For target sequences anchored to nitrocellulose or nylon-66 (e.g. GeneScreen, Nytran, or the like), prehybridization treatment can comprise treatment with 5-10x Denhardt's solution, with 2-6x SSC preferably containing a mild detergent, e.g. 0.5% sodium dodecylsulfate (SDS), for 15 min. to 1 hr. at a temperature in the range of about 25° to 60°. Denhardt's solution, disclosed in Biochem. Biophys. Res. Commun., Vol. 23, pgs. 641-645 (1966), consists at 10x concentration of 0.2% bovine serum albumin, 0.2% polyvinylpyrolidone, and 0.2% Ficoll. SSC, another standard

reagent in the hybridization art, consists at 1x of 0.15 M NaCl, 0.015 M sodium citrate, at pH 7.0. Preferred treatment times, temperatures, and formulations may vary with the particular embodiment.

Preferably, the probe is applied to the anchored DNA at a concentration in the range of about 1-10 ng/ml in a solution substantially the same as the prehybridization solution, e.g. 5-10x Denhardt's solution with 2-6x SSC and a mild detergent, e.g. 0.5% SDS. More preferably, the probe concentration is in the range of about 1-5 ng/ml. Preferably, the hybridization is carried out at a temperature 10-20^{o} C below the expected 50% dissociation temperatue, T_d, between the probe and the target. That is, a temperature is selected at which a high proportion, e.g. greater than 80-90%, of all the perfectly matched probes form stable duplexes. For 8-mer probes the preferred hybridization temperature is in the range of about 10-18^{o} C. Hybridization times are preferably in the range of about 3-16 hours. Different hybridization times may be selected for probes of different degrees of degeneracy, because the effective consentrations of particular sequences within a highly degenerate probe, e.g. 0A000000, are considerably less than those of particular sequences in a low degeneracy probe, e.g. AAA0AAAA. Thus, higher Cot values (which usually means longer hybridization times) may be required for more degenerate probes to attain a sufficient degree of binding of perfectly matched sequences. Preferably, a single hybridization time is selected for all probes which is determined by the hybridization kinetics of the most degenerate probe. Probe degeneracy becomes important when relative signals are compared after autoradiography. Probes having higher degeneracy will produce lower signals than probes of lower degeneracy. Therefore, relative signals should only be compared among autoradiographs associated with probes of equivalent degeneracy. Signal comparisons are aided by the simultaneous running of positive and negative controls for each probe, or at least for each degeneracy class.

Removing nonspecifically bound and mismatched probe sequences by washing is an important aspect of the invention. Temperatures and wash times are selected which permit the removal of a maximum amount of nonspecifically bound and mismatched probe sequences, while at the same time permit the retention of a maximum number of probe sequences forming perfectly matched duplexes with the target. The length and base composition of the probe are two important factors in determining the appropriate wash temperature. For probe lengths in the preferred range of 7-11 bases, the difference in duplex stability between perfectly matched and mismatched probe sequences is quite large, the respective T_d's differing by perhaps as much as 10^{o} C, or more. Thus, it is not difficult to preferentially remove mismatches by adjusting wash temperature. On the other hand, composition differences, e.g. G-C content versus A-T content, give rise to a broadened range of probe T_d's, and lower wash temperatures reduce the ability to remove nonspecifically bound probe. Consequently, the wash temperature must be maximized for removing nonspecifically bound probe, yet it cannot be so high as to preferentially remove perfectly matched probes with relatively high A-T content.

Preferably, hybridization conditions and/or nucleotide analogs are selected which minimize the difference in binding energies of the various base pairs, in order to minimize sequence-specific differences in probe binding. Such minimization is preferable because it increases the sensitivity with which perfectly matched probes can be detected. Sensitivity is increased because such minimization makes the transition from probe/target duplexes to single stranded probe and single stranded target much sharper when temperature is increased, i.e. the probe/target T_m is less broad. For example, when hybridization occurs in the presence of tetraalkylammonium salts, the differences in binding energy between G-C pairs and A-T pairs is reduced, e.g. Wood et al, <u>Proc. Natl. Acad. Sci.</u>, Vol. 82, pgs. 1585-1588 (1985). Likewise, use of alpha-anomeric nucleoside analogs results in

stronger binding energies, e.g. Moran et al, Nucleic Acids Research, Vol. 16, pgs. 833-847 (1988); and the use of 2-aminoadenine in place of adenine results in stronger binding, e.g. Chollet et al, Nucleic Acid Research, Vol 16, pgs. 305-317 (1988). Accordingly, a preferred wash procedure for 8-mer probes comprises washing the filters three times for about 15 minutes in 6x SSC containing 0.5% SDS at a temperature in the range of about 10-12^{o} C, followed by one or two rinses with 3.0 M Me_4NCl, 50 mM Tris-HCl, pH 8.0, 0.5% SDS, at a temperature in the range of about 10-12^{o} C, followed by a 1.5-2.5 minute wash in 3.0 M Me_4NCl, 50 mM Tris-HCl, pH 8.0, 0.5% SDS, at a temperature in the range of about 24-28^{o} C. For 9-mer probes, the procedure is substantially the same, except that the final wash temperature is preferably in the range of about 26-30^{o}C. After hybridization and washing, quantitative measurements of bound probe are carried out using standard techniques, e.g. for radiolabelled probes, autoradiography or scintillation counting can be used.

III. Sequence Reconstruction

The general nature of the reconstruction problem is illustrated by the example of Figure 1, in which four subsets of 4-mer probes are used to analyze the sequence of the 21-mer, CGAATGGAACTACCGTAACCT. On the left of Figure 1 is a list of 4-mer probes having every possible permutation of fixed and non-fixed positions with respect to deoxyadenosine, deoxycytosine, deoxyguanosine, and thymidine, respectively, for the following combinations of fixed bases and non-fixed bases: 1 fixed and 3 non-fixed, 2 fixed and 2 non-fixed, 3 fixed and 1 non-fixed, and 4 fixed and 0 non-fixed. That is, the list contains at least one probe having a sequence of fixed and non-fixed positions with respect to A, C, G, and T equivalent to every possible permutation of A's and non-fixed positions, C's and non-fixed positions, G's and non-fixed positions, and T's and non-fixed positions, respectively. In the figure, there is one probe for each row of a two dimensional array having a number of columns equal to the

length of the unknown sequence, in this example 21. The data obtained by separately hybridizing the 60 probes to the 21-mer are listed under the column, "Perfect Matches." The data represent the number of each probe type having perfect complementarity with a four base subsequence of the 21- mer. Under the 21-mer sequence itself the probes are positioned along the sequence where perfect complementarity occurs. The objective of a reconstruction algorithm is to determine the positions of enough probes so that the target sequence can be reconstructed.

The reconstruction problem can be approached in many ways. The problem is related to the traveling salesman problem in that it involves finding a permutation of objects which is in some sense optimal. There is an extensive literature on such combinatorial problems which provides guidance in formulating the best approach for a particular embodiment, e.g. Lawler et al, eds., The Traveling Salesman Problem: A Guided Tour of Combinatorial Optimization (John Wiley & Sons, New York, 1985); Kirkpatrick, J. Stat. Phys., Vol. 34, pgs. 975-986 (1984); Held and Karp, J. Soc. Indust. Appl. Math., Vol. 10, pgs. 196-210 (1962); and Lin and Kernighan, Oper. Res., Vol 21, pgs. 498-516 (1973).
A preferred approach to the reconstruction problem requires that the target sequence include one or more known sequence regions. In particular, a first known sequence region is located at one end of the target sequence and a second known sequence region is located at the other end of the target sequence. The presence of the two known sequence regions permits the construction of a simplified and efficient reconstruction algorithm. Roughly, the reconstruction problem is a problem of finding an ordering of overlapping probe sequences which corresponds to the target sequence. The known sequence regions define the starting and ending probes sequences in a reconstruction. The intervening unknown sequence region can be reconstructed from the remaining probe sequences by requiring that each successively selected probe properly overlap the previously selected probe sequence.

Figure 2 is a flow chart of such an algorithm. It consists of two parts which are performed alternatively, drawing probes from the same set (referred to herein as the data set) determined by the hybridization data: (1) construction of candidate sequences from properly overlapping fixed-initial-position and fixed-final position probes starting from one of the known sequence regions, and (2) construction of candidate sequences from properly overlapping fixed-final-position and fixed-initial-position probes starting from the other known sequence region. The term "properly overlapping" simply means overlapping in the sense described at the beginning of this section and illustrated in Figure 1. Thus, in this algorithm only probes having either a fixed initial position (3') or a fixed final position (5'), or both, are employed in the reconstruction. These two classes of probes are referred to herein as FIP probes and FFP probes.

For the algorithm, two sets of numbers (or logical variables depending on the implementation) are defined by the nucleotide sequences of the first and second known sequence regions. These sets are referred to as the initial left register and the initial right register **2**, respectively. The size of the registers depends on the length of the probes employed. Usually the registers have L-1 elements, or entries, where L is the length of the probe. Starting with the initial right register, the algorithm compares the entries of the register with every FFP and FIP probe that forms a perfect match with the target sequence, **4**. The comparison is between bases 2 through L of the selected probe and the numbers (or entries) 1 through L-1 of the right register. That is, base at position 2 is compared to the entry at position 1 of the right register, base at position 3 is compared to the entry at position 2 of the right register, and so on. Initially, as stated above, the entries of the registers are determined by the bases of the first and second known sequence regions. If the comparison results in proper overlap in each of the L-1 positions, then the current contents of the register are loaded into a new right register

and then the entries of the new right register are shifted to the right one position. That is, entry 1 of the new right register is moved to position 2, entry 2 is moved to position 3, and so on. Entry L-1 is discarded, and the fixed base at the initial position of the probe (or some representation of it) is loaded into position 1. Next, for the new right register and selected probe to be retained for further comparisons, an FFP probe must be found that properly overlaps the register and the initial fixed base of the selected FIP probe (unless of course the FIP probe is also an FFP probe) **8**. When such selections are made (i.e. **4** and **8**) the selected probe(s) are removed from the data set. The new right register is saved along with and associated set of properly overlapping FIP probes whose selections led to the current register, and an associated set of properly overlapping FFP probes.

After each new right register is formed, one or more left registers are formed, **16** and **18**, by extending preexisting left registers in substantially the same way as the right registers, excepts that FFP probes are selected first and positions 1-L of the FFP probe are compared to entries 1-L of the left register. The FIP and FFP probes are selected from the probes remaining in the data set. That is, any probes previously selected to "extend" the right or left registers cannot be selected. This also holds for right registers formed in successive iterations after the first. As a result of these comparisons, pairs of right and left registers are formed, and associated with each pair are four sets of probes, **20**: (i) the set of FIP probes selected to extend the right register, (ii) the set of FFP probes selected to properly overlap the right register and FIP probe, (iii) the set of FFP probes selected to extend the left register, and (iv) the set of FIP probes selected to properly overlap the left register and FFP probe. At each step i (see Figure 2), M_{i+1} such pairs and associated sets are formed.

The comparisons between probe and register positions are carried out as follows. The register entries are always the

bases A, C, G, or T (or some representation thereof). The probe positions are always occupied by a base or the absence of a base. Recall from above that probes can be represented by the notation, for example, A0AA000A. The 0's represent in this case represent either C, G, T, or a degeneracy-reducing analog thereof. In other words, the 0's represent "not A's". The comparisons entail the determination of the truth value of a base (from the register) and a base or a negative of a base (from the probe being compared). For example, if the register entry is A and the probe entry is "not T", then the logical operation of "A AND not T" is logically true. Thus, proper overlap exists. On the other hand, is the probe entry is "not A", then the logical operation of "A AND not A" is logically false. Thus, the overlap is improper and the probe is rejected.

In successive steps, each of the M_i pairs of registers are compared to probes of their respective data sets, generating in turn, a set of M_{i+1} pairs of registers. With each step the data set is reduced in size by two or more probes, and the respective canditate sequences are increased in size by one base each. When a register is compared to each of the remaining probes in its associated data set and no probe is found that properly overlaps, the register and its associated sets are discarded **6-14**. The algorithm halts when every probe in the data sets have been used (i.e., sorted into one of the four associated sets). If more than one candidate sequences are generated, or if it is desired to check the consistency of the data, the same process of repeated rounds of comparisons can be carried out starting with the initial left register and set of probes that have fixed final positions <u>and</u> a perfect match with the target sequence **10-16**. In this case, entries 1 through L-1 of the initial left register are compared with probe positions 1 through L-1, respectively, and successive registers, R_j after the jth round of comparisons, are generated by shifting current entries to the left and entering the final fixed base of the properly overlapping probe to position L-1 of the new

register. In a similar manner to that described above, additional candidate sequences are generated, and are compared to the ones previously generated. Only ones that occur in both sets are retained 18. Further eliminations are possible by requiring that all of the remaining non-fixed final and non-fixed initial position probes find properly overlapping positions within each of the candidates 20.

Preferably, the algorithm is implemented on a computer with parallel processing capabilities. For example, the algoithm of appendix I can be loaded onto each of the nodes of a hypercube parallel processor, e.g. a 1024 node NCUBE/ten computer (Ncube Corp., Beaverton, Oregon).

The above algorithm does not necessarily give a unique solution in every case. Generally, regions of high frequency repeats (e.g. ACACACAC..., GTCGTCGTC..., or the like) or constant regions (e.g. AAAAAAA..., or the like) substantially longer than the probe give rise to non-unique solutions. For example, it is impossible to uniquely reconstruct (with the above algoritm) target sequences which contain long stretches of a single base type within which a few bases of a different type are clustered. Thus, if 4-mer probes were used to reconstruct a sequence with a stretch containing --AAAAAAAAAAAAAAAAAAAAAGCATAAAAAAAAAAAAAAA--the position of GCAT within the sequence of A's cannot be unequivocally determined. In some cases, if alternative solutions are found the correct sequence can be discerned by sequencing the non-uniquely determined portions of the target sequence by standard techniques.

Example I. <u>Sequence Determination of the 119 Basepair Sca-Xmn Fragment of pUC19 with 8-mer Probes</u>

A 119 basepair double stranded DNA is obtained by <u>Xmn</u> I and <u>Sca</u> I restriction endonuclease digestion of the pUC19 plasmid, described by Yanisch-Perron et al, <u>Gene</u>, Vol. 33, pgs. 103-119 (1985), and widely available commercially, e.g. Bethesda Research Laboratories (Gaithersburg, MD). Large scale isolation of pUC19 can be carried out by standard

procedures, e.g. as disclosed by Maniatis et al, Molecular Cloning: A Laboratory Manual, (Cold Spring Harbor Laboratory, New York, 1982) (i.e. alkali lysis followed by equilibrium centrifugation in cesium chloride-ethidium bromide gradient). Alternatively, purified pUC19 is purchased commercially as needed, e.g. from Bethesda Research Laboratories.

1 mg pUC19 DNA is precipitated with 95% ethanol, dried, and resuspended in 1 ml of Sca-Xmn restriction buffer (e.g., 50 mM NaCl, 10 mM Tris-HCl (pH 7.8), 10 mM $MgCl_2$, 10mM 2-mercaptoethanol, 100 ug bovine serum albumin) for about 2.0 hours at 37°C. After stopping the reaction by adding 0.5 M EDTA (pH 7.5), the restriction buffer is mixed with xylene cyanol and loaded onto a 8 percent polyacrylamide gel for electrophoresis. The band containing the 119 basepair fragment is excised, and the DNA eluted as described by Maniatis et al, page 178 (cited above).

The fragments are resuspended at 1 ng/100 ul of 0.2 N NaOH for 10 minutes, chilled, and mixed with an equal volume of 10xSSC (1.5 M sodium chloride and 0.15 M sodium citrate). 100 ul samples of this fragment solution are pipetted into the wells of slot-blotting apparatus, e.g. eleven 72-well Minifold II micro-sample filtration manifolds, available from Schleicher and Scheull, Keene, NH), each apparatus holding a GeneScreen membrane that had been previously wetted for 15-20 minutes in 1xSSC. After 2 hours, the solution is gently sucked through the membranes, washed with 2xSSC, and allowed to dry. After drying, the membranes are baked at 80°C for 2-4 hours. Before application of probe, the membranes are treated with prehybridization mixture (10x Denhardt's with 0.5% SDS for 1 hour at 60°C, followed by washing with 2xSSC).

Probes for hybridization are synthesized by phosphoramidite chemistry on an Applied Biosystems, Inc. model 380A DNA synthesizer. 4 X 196=784 mixed oligonucleotide probes are employed, a probe for each kind of 8-mer sequence having either a fixed initial position or a fixed final position (see Appendix I). Non-fixed positions of the cytosine

and adenosine subsets of probes are filled by deoxyadenosine and deoxyinosine and deoxycytosine and deoxyinosine, respectively. The probes are ^{32}P labelled following the T4 polynucleotide kinase protocol of Maxum and Gilbert (Meth Enzymol., Vol. 65, pgs. 497-560 (1980)), applied to the manifold wells at 18°C for 16 hours at a concentration of 1 ng/ml in 500 ul of hybridization mixture consisting of 5x Denhardt's, 5xSSPE, and 0.5% SDS. After hybridization the membranes are washed 3 times with 6xSSC containing 0.5% SDS at 12°C, followed by 2 rinses with 3.0 M Me_4NCl, 50 mM Tris-HCl (pH 8.0), 0.5% SDS at 12°C, and a final 2.0 minute wash in 3.0 M Me_4NCl, 50 mM Tris-HCl (pH 8.0), 0.5% SDS at 26-27°. After washing, the dried membranes are autoradiographed on XAR-5 film (or its equivalent) for 2-4 days. "Slots" on the developed film are analyzed on a LKB UltroScan XL Laser Densitometer, or like instrument.

Numbers of perfectly matched probes are determined by comparing the relative signal strengths of probes having the same degree of degeneracy. Also, because a double stranded target sequence is used, the values for probe number used in the reconstruction algorithm are the average of the signal for each probe type and its complement (with respect to the fixed bases). The sequence is reconstructed from the probe number data by program RCON8, whose source code is listed in the Appendix. RCON8 assumes that the eight base sequences on each end of the target sequence are known sequence regions. The program returns the noncoding sequence listed in a 3'-5' orientation from left to right.

Example II. Sequence Determination of the 323 Basepair Pvu II Fragment of pUC19 Using 9-mer Probes

A 323 basepair double stranded DNA is one of two fragments obtained by Pvu II digestion of pUC19. The same procedure is followed as described in Example I for preparing the 323 basepair Pvu II fragments, denaturing them, and anchoring them to GeneScreen substrates. Pre-hybridization, hybridization, and wash protocols are the same, except that

the final high temperature wash is carried out at 28-29^{o}C.

Probes are synthesized and labeled as in Example I. 1556 probes are employed. A probe is prepared for each 9-mer sequence having a fixed base at the initial or final position (i.e. 389 each for A, C, G, and T probes). As in Example I, probes of the form 100000000 and 000000001 are replaced by three probes having different types of fixed bases, e.g. 00000000T is replaced by A0000000T, C0000000T, and G0000000T.

The sequence of the Pvu fragment is reconstructed with a modified version of the program of Appendix I which specifically accommodates 9-mer probe data. Like the 8-mer version, the program assumes that the nine base sequences on each end of the target sequence are known sequence regions.

Brief Description of the Drawings

Figure 1 illustrates the general problem of sequence reconstruction by showing how a 21-mer sequence can be reconstructed by four subsets of 4-mer probes.

Figure 2 is a flow chart diagrammatically illustrating a preferred reconstruction algorithm.

Appendix I. Source Code Listing of Reconstruction Algorithm RCON8

```
      program RCON8
c
c            Program RCON8  reconstructs sequences from 8-mer
c              probes having fixed bases at their initial
c              positions and fixed bases at their final
c              positions.
c
      implicit integer*2 (a-z)
      dimension rl(xx,120),r2(xx,120),ll(xx,120),l2(xx,120)
      dimension dr2(xx,120),dll(xx,120),dl2(xx,120),drlnum(xx)
      dimension dr2num(xx),dllnum(xx),dl2num(xx),pset(240,2)
      dimension pnum8(4,259,8),rregl(xx,7),rreg2(xx,7)
      dimension lreg2(xx,7),lregl(xx,7)
      dimension tab(4,-4:4),drl(xx,120)
      dimension na(259),nc(259),ng(259),nt(259)
      character*1 seql(120),seqr(120),s,pseq8(4,259,8)
c
      character*10 pdata
      common rl,r2,ll,l2,drl,dr2,dll,dl2,rregl,rreg2,lregl,
     1       pnum8,tab,drlnum,dr2num,dllnum,dl2num,npx,np0,
     2       lreg2,pset,pseq8
c
c      READ pseq8 (list of all possible probe types) from
c              data file.
c      GENERATE pnum8 from pseq8.
c      READ probe data from data file and load into arrays
c              na, nc, ng, and nt.
c      GENERATE pset from pseq and na, nc, ng, and nt.
c      READ register transition table values from data file
c              and load into array tab.
c      ENTER bases of known sequence regions into arrays
c              rreg and lreg.
c
      numreg=1
c
c         NUMREG is the current number of rregisters
c         NPX is the number of probes in the data having
c             an initial (3') fixed base.
c         NPXX is the number of probes in the data having
c             both an initial (3') and final (5') fixed base.
c
      halt=int((npxx + (npx-npxx)/2)/2)
      rl(1,1)=0
      ll(1,1)=0
      drlnum(1)=0
```

```
        dllnum(1)=0
         ii=0
1000     ii=ii+1
c
c          ii indexes the round of comparisons.  ii is also
c          equal to the current length of candidate sequences.
c
         w=0
         do 1100 f=1,numreg
            do 1200 kk=1,npx-np0
               if(ii.eq.1) then
                if(kk.gt.1 .and. pset(kk,1).eq.pset(r2(w,ii),1)
     1               .and. pset(kk,2).eq.pset(r2(w,ii),2) .and.
     2               skipa.eq.1) goto 1200
                  skipa=0
                  do 1550 j=1,7
1550              if(tab(rregl(f,j),pnum8(pset(kk,2),pset(kk,1),
     1                     j+1)).eq.0) goto 1200
               else
c
               do 1300 mm=1,ii-1
1300            if(kk.eq.rl(f,mm) .or. kk.eq.ll(f,mm)) goto 1200
                  do 1400 mm=1,dllnum(f)
1400                 if(kk.eq.dll(f,mm)) goto 1200
                  if(kk.gt.1 .and. pset(kk,1).eq.pset(r2(w,ii),1)
     1               .and. pset(kk,2).eq.pset(r2(w,ii),2) .and.
     2               skipa.eq.1) goto 1200
                  skipa=0
               do 1500 j=1,7
1500              if(tab(rregl(f,j),pnum8(pset(kk,2),pset(kk,1),
     1               j+1)).eq.0) goto 1200
              endif
c
            skipb=0
              if(pnum8(pset(kk,2),pset(kk,1),8).lt.0 .or.
     1           pset(kk,1).le.3) then
                 do 1600 jj=npx-np0+1,npx
                  if(ii.eq.1) then
           if(jj.gt.1 .and. pset(jj,1).eq.pset(dr2(w,dr2num(w)),
     1           1) .and. pset(jj,2).eq.pset(dr2(w,dr2num(w)),2)
     2                 .and. skipb.eq.1) goto 1600
                    skipb=0
                       do 1950 x=1,8
                       if(x.eq.1) then
                   if(tab(pset(kk,2),pnum8(pset(jj,2),pset(jj,1),
     1                           1)).eq.0) goto 1600
                       else
                          if(tab(rregl(f,x-1),pnum8(pset(jj,2),
```

```
     1                    pset(jj,1),x)).eq.0) goto 1600
                     endif
 1950                continue
               else
                  do 1700 mm=1,drlnum(f)
 1700                if(jj.eq.drl(f,mm)) goto 1600
                  do 1800 mm=1,ii-1
 1800                if(jj.eq.ll(f,mm)) goto 1600
         if(jj.gt.1 .and. pset(jj,1).eq.pset(dr2(w,dr2num(w)),
      1        1) .and. pset(jj,2).eq.pset(dr2(w,dr2num(w)),2)
      2              .and. skipb.eq.1) goto 1600
                  skipb=0
                  do 1900 x=1,8
                     if(x.eq.1) then
                if(tab(pset(kk,2),pnum8(pset(jj,2),pset(jj,1),
      1                    1)).eq.0) goto 1600
                     else
                        if(tab(rregl(f,x-1),pnum8(pset(jj,2),
      1                    pset(jj,1),x)).eq.0) goto 1600
                     endif
 1900                continue
               endif
                  ksave=kk
                  jsave=jj
                  w0=w
 c
                  call left(ii,w,f,jsave,ksave)
 c
                  if(w.eq.w0) goto 1600
                  do 2000 k=w0+1,w
                     do 2100 i=1,6
 2100                   rreg2(k,i+1)=rregl(f,i)
                        rreg2(k,1)=pset(kk,2)
                     do 2600 q=1,drlnum(f)
 2600                   dr2(k,q)=drl(f,q)
                     dr2num(k)=drlnum(f) + 1
                     dr2(k,dr2num(k))=jj
 2000                continue
                  if(ii.eq.1) then
                       do 2200 k=w0+1,w
 2200                     r2(k,1)=kk
                  else
                       do 2300 k=w0+1,w
                          do 2400 x=1,ii-1
 2400                        r2(k,x)=rl(f,x)
 2300                        r2(k,ii)=kk
                  endif
                  skipa=1
```

```
                skipb=l
1600            continue
c
              else
c
                jsave=0
                ksave=kk
                w0=w
c
                call left(ii,w,f,jsave,ksave)
c
                if(w.eq.w0) goto 1200
                do 2700 k=w0+1,w
                   do 2800i=1,6
2800                  rreg2(k,i+1)=rregl(f,i)
                      rreg2(k,l)=pset(kk,2)
                   do 2900 q=l,drlnum(f)
2900                  dr2(k,q)=drl(f,q)
                   dr2num(k)=dr2num(f)
2700               continue
                if(ii.eq.l) then
                   do 6000 k=w0+1,w
6000                  r2(k,l)=kk
                else
                   do 6100 k=w0+1,w
                      do 6200 x=l,ii-1
6200                     r2(k,x)=rl(f,x)
6100                     r2(k,ii)=kk
                endif
                skipa=l
              endif
1200          continue
1100      continue
c
              numreg=w
                    do 7000 k=l,numreg
                       do 7100 m=l,ii
                          rl(k,m)=r2(k,m)
7100                      ll(k,m)=l2(k,m)
                       do 7200 m=l,7
                          rregl(k,m)=rreg2(k,m)
7200                      lregl(k,m)=lreg2(k,m)
                       do 7300 m=l,dr2num(k)
7300                      drl(k,m)=dr2(k,m)
                       do 7400 m=l,dl2num(k)
7400                      dll(k,m)=dl2(k,m)
                       drlnum(k)=dr2num(k)
                       dllnum(k)=dl2num(k)
```

```
7000              continue
           if (ii.lt.halt) goto 1000
c
c                 PRINT SEQUENCES
c
3000       continue
5000     end
c
c
         subroutine left(ii,w,f,jsave,ksave)
         implicit integer*2 (a-z)
         dimension rl(xx,120),r2(xx,120),l1(xx,120),l2(xx,120)
         dimension dr2(xx,120),dl1(xx,120),dl2(xx,120),drlnum(xx)
         dimension dr2num(xx),dl1num(xx),dl2num(xx),pset(240,2)
         dimension pnum8(4,259,8),rregl(xx,7),rreg2(xx,7)
         dimension lreg2(xx,7),lregl(xx,7)
         dimension tab(4,-4:4),drl(xx,120)
         character*1 pseq8(4,259,8)
         common rl,r2,l1,l2,drl,dr2,dl1,dl2,rregl,rreg2,lregl,
      1        pnum8,tab,drlnum,dr2num,dl1num,dl2num,npx,np0,
      2        lreg2,pset,pseq8
c
            skipa=0
         do 1000 hh=1,npx
            if(pnum8(pset(hh,2),pset(hh,1),8).lt.0 .or.
      1        pset(hh,1).le.3 .or. hh.eq.jsave) goto 1000
            if(ii.eq.1) then
               if(hh.gt.1 .and. pset(hh,1).eq.pset(l2(w,ii),1)
      1           .and. pset(hh,2).eq.pset(l2(w,ii),2) .and.
      2           skipa.eq.1) goto 1000
              skipa=0
                do 1350 j=1,7
1350           if(tab(lregl(f,j),pnum8(pset(hh,2),pset(hh,1),j))
      1                .eq.0) goto 1000
            else
c
            do 1100 mm=1,ii-1
1100           if(hh.eq.rl(f,mm) .or. hh.eq.l1(f,mm)) goto 1000
               do 1200 mm=1,drlnum(f)
1200              if(hh.eq.drl(f,mm)) goto 1000
               if(hh.gt.1 .and. pset(hh,1).eq.pset(l2(w,ii),1)
      1           .and. pset(hh,2).eq.pset(l2(w,ii),2) .and.
      2           skipa.eq.1) goto 1000
              skipa=0
              do 1300 j=1,7
1300           if(tab(lregl(f,j),pnum8(pset(hh,2),pset(hh,1),j))
      1             .eq.0) goto 1000
            endif
```

```
c
               skipb=0
              if(pset(hh,1).ge.137) then
                  do 1400 rr=1,npx-np0
                  if(rr.eq.ksave) goto 1400
                  if(ii.eq.1) then
            if(rr.gt.1 .and. pset(rr,1).eq.pset(dl2(w,dl2num(w)),
     1            1) .and. pset(rr,2).eq.pset(dl2(w,dl2num(w)),2)
     2                  .and. skipb.eq.1) goto 1400
                     skipb=0
                    do 1750 x=1,8
                        if(x.eq.8) then
                           if(tab(pnum8(pset(hh,2),pset(hh,1),x),
     1                      pnum8(pset(rr,2),pset(rr,1),x)).eq.0)
     2                        goto 1400
                        else
                    if(tab(lregl(f,x),pnum8(pset(rr,2),pset(rr,1),
     1                        x)).eq.0) goto 1400
                        endif
1750                 continue
                 else
c
                     do 1500 mm=1,dllnum(f)
1500                    if(rr.eq.dll(f,mm)) goto 1400
                     do 1600 mm=1,ii-1
1600                    if(rr.eq.rl(f,mm)) goto 1400
             if(rr.gt.1 .and. pset(rr,1).eq.pset(dl2(w,dl2num(w)),
     1             1) .and. pset(rr,2).eq.pset(dl2(w,dl2num(w)),2)
     2                 .and. skipb.eq.1) goto 1400
                     skipb=0
                     do 1700 x=1,8
                        if(x.eq.8) then
                           if(tab(pnum8(pset(hh,2),pset(hh,1),x),
     1                       pnum8(pset(rr,2),pset(rr,1),x)).eq.0)
     2                        goto 1400
                        else
                    if(tab(lregl(f,x),pnum8(pset(rr,2),pset(rr,1),
     1                        x)).eq.0) goto 1400
                        endif
1700                 continue
                 endif
                     skipa=1
                     skipb=1
                     w=w+1
                     dl2num(w)=dllnum(f)+1
                     do 1710 k=1,dllnum(f)
1710                    dl2(w,k)=dll(f,k)
                        dl2(w,dl2num(w))=rr
```

```
c
                  do 1800 i=1,6
1800                 lreg2(w,i)=lreg1(f,i+1)
                     lreg2(w,7)=pset(hh,2)
                  if(ii.eq.1) then
                      l2(w,1)=hh
                  else
                      do 1900 k=1,ii-1
1900                     l2(w,k)=l1(f,k)
                         l2(w,ii)=hh
                  endif
1400          continue
            else
               skipa=1
               w=w+1
               dl2num(w)=dl1num(f)
               do 2200 j=1,dl1num(f)
2200              dl2(w,j)=dl1(f,j)
               do 2000 i=1,6
2000              lreg2(w,i)=lreg1(f,i+1)
                  lreg2(w,7)=pset(hh,2)
               if(ii.eq.1) then
                   l2(w,1)=hh
               else
                   do 2100 k=1,ii-1
2100                  l2(w,k)=l1(f,k)
                      l2(w,ii)=hh
               endif
            endif
1000     continue
         return
         end
```

I CLAIM:

1. A method for determining the nucleotide sequence of a nucleic acid, the method comprising the steps of:

providing a set of probes, each probe within the set having a predetermined length and each probe within the set having a predetermined sequence of fixed and non-fixed positions, the fixed positions comprising one or more predetermined kinds of nucleotides;

hybridizing the probes of the set to the nucleic acid;

determining the number of copies of each probe in the set that form perfectly matched duplexes with the nucleic acid; and

reconstructing the nucleotide sequence of the nucleic acid from the predetermined sequences of the probes that form perfectly matched duplexes with the nucleic acid.

2. The method of claim 1 wherein said set contains at least one probe comprising a sequence of fixed and non-fixed positions equivalent to that of each permutation of a plurality of fixed and non-fixed positions equal to or less than the length of the probe.

3. The method of claim 2 further including the steps of:

anchoring a known quantity of said nucleic acid to each of a plurality of solid phase supports; and washing each of the solid phase supports after hybridizing said probes so that substantially all of said probes not forming perfectly matched duplexes with said nucleic acid are removed from the solid phase support, and so that substantially all of said probes forming perfectly matched duplexes with said nucleic acid remain on the solid phase support.

4. The method of claim 3 wherein said step of hybridizing includes separately hybridizing each probe of said set to said nucleic acid on a different solid phase support of said plurality of solid phase supports.

5. The method of claim 4 wherein said predetermined length of said probes are in the range of from seven to eleven nucleotides, inclusive.

6. The method of claim 5 wherein said non-fixed positions of said probes are occupied by at least one degeneracy-reducing analog.

7. A method for determining the nucleotide sequence of a nucleic acid, the method comprising the steps of:

providing a first set of probes, each probe within the first set having the same length, the length being from seven to ten nucleotides, and each probe within the first set having a predetermined sequence of fixed and non-fixed bases, the fixed bases being deoxyadenosine and the non-fixed bases comprising deoxycytosine, deoxygauanosine, thymidine, or a degeneracy-reducing analog thereof, such that for each permutation of fixed and non-fixed bases less than or equal to the length of the probe, the first set contains at least one probe having a sequence equivalent to such permutation;

providing a second set of probes, each probe within the second set having the same length, the length being from seven to ten nucleotides, and each probe within the second set having a predetermined sequence of fixed and non-fixed bases, the fixed bases being deoxycytosine and the non-fixed bases comprising deoxyadenosine, deoxygauanosine, thymidine, or a degeneracy-reducing analog thereof, such that for each permutation of fixed and non-fixed bases less than or

equal to the length of the probe, the second set contains at least one probe having a sequence equivalent to such permutation;

providing a third set of probes, each probe within the third set having the same length, the length being from seven to ten nucleotides, and each probe within the third set having a predetermined sequence of fixed and non-fixed bases, the fixed bases being deoxyguanosine and the non-fixed bases comprising deoxyadenosine, deoxycytosine, thymidine, or a degeneracy-reducing analog thereof, such that for each permutation of fixed and non-fixed bases less than or equal to the length of the probe, the third set contains at least one probe having a sequence equivalent to such permutation;

providing a fourth set of probes, each probe within the fourth set having the same length, the length being from seven to ten nucleotides, and each probe within the fourth set having a predetermined sequence of fixed and non-fixed bases, the fixed bases being thymidine and the non-fixed bases comprising deoxyadenosine, deoxycytosine, deoxyguanosine, or a degeneracy-reducing analog thereof, such that for each permutation of fixed and non-fixed bases the length of the probe, the fourth set contains at least one probe having a sequence equivalent to such permutation;

anchoring a known quantity of the nucleic acid to each of a plurality of solid phase supports;

separately hybridizing each probe of the first, second, third, and fourth sets to the nucleic acid anchored on the solid phase supports;

washing each of the solid phase supports after hybridizing said probes so that substantially all of said probes not forming perfectly matched duplexes with the nucleic acid are removed from the solid phase support, and so that substantially all of said probes forming perfectly matched duplexes with the nucleic

acid remain on the solid phase support;

determining the number of copies of each probe in each set that form perfectly matched duplexes with the nucleic acid; and

reconstructing the nucleotide sequence of the nucleic acid from the predetermined sequences of the probes that form perfectly matched duplexes with the nucleic acid.

8. The method of claim 7 wherein said nucleic acid contains at least one known sequence region.

9. The method of claim 8 wherein said probe of said first set having said length from eight to nine nucleotides, said probe of said second set having said length from eight to nine nucleotides, said probe of said third set having said length from eight to nine nucleotides, and said probe of said fourth set having said length from eight to nine nucleotides.

10. The method of claim 9 wherein said degeneracy-reducing analog of said first set includes deoxyinosine, 5-fluorodeoxyuridine, and N^4-methoxycytosine, said degeneracy-reducing analog of said second set includes deoxyinosine and 2-aminopurine, and said degeneracy-reducing analog of said third set includes 2-aminopurine and N^4-methoxycytosine.

11. The method of claim 10 wherein said step of washing includes exposing said solid phase support to tetramethylammonium chloride at a concentration of between about 2 to 4 moles per liter.

TARGET SEQUENCE

PROBE	5'-C	G	A	A	T	G	G	A	A	C	T	A	C	C	G	T	A	A	C	C	T-3'	PERFECT MATCHES
3'A000					A	0	0	0			A	0	0	0		A	0	0	0			3
0A00				0	A	0	0			0	A	0	0		0	A	0	0				3
00A0			0	0	A	0			0	0	A	0		0	0	A	0					3
000A		0	0	0	A			0	0	0	A		0	0	0	A		0	0	0	A	4
AA00																						0
A0A0																						0
A00A																						0
0AA0																						0
0A0A																						0
00AA																						0
AAA0																						0
AA0A																						0
A0AA																						0
0AAA																						0
AAAA																						0
C000		C	0	0	0		C	0	0	0					C	0	0	0				3
0C00	0	C	0	0										0	C	0	0					2
00C0													0	0	C	0						1
000C												0	0	0	C							1
CC00						C	C	0	0													1
C0C0																						0
C00C																						0
0CC0					0	C	C	0														1
0C0C																						0
00CC				0	0	C	C															1
CCC0																						0
CC0C																						0
C0CC																						0
0CCC																						0
CCCC																						0
G000	G	0	0	0									G	0	0	0						2
0G00									0	G	0	0										1
00G0								0	0	G	0											1
000G							0	0	0	G						0	0	0	G			2
GG00													G	G	0	0						1
G0G0																						0
G00G										G	0	0	G									1
0GG0												0	G	G	0			0	G	G	0	2
0G0G																						0
00GG											0	0	G	G			0	0	G	G		2
GGG0																						0
GG0G																						0
G0GG																						0
0GGG																						0
GGGG																						0
T000				T	0	0	0					T	0	0	0			T	0	0	0	3
0T00											0	T	0	0								1
00T0										0	0	T	0									1
000T					0	0	0	T						0	0	0	T					2
TT00			T	T	0	0		T	T	0	0						T	T	0	0		3
T0T0																						0
T00T									T	0	0	T										1
0TT0		0	T	T	0		0	T	T	0						0	T	T	0			3
0T0T																						0
00TT	0	0	T	T		0	0	T	T						0	0	T	T				3
TTT0																						0
TT0T																						0
T0TT																						0
0TTT																						0
TTTT																						0

FIGURE 1

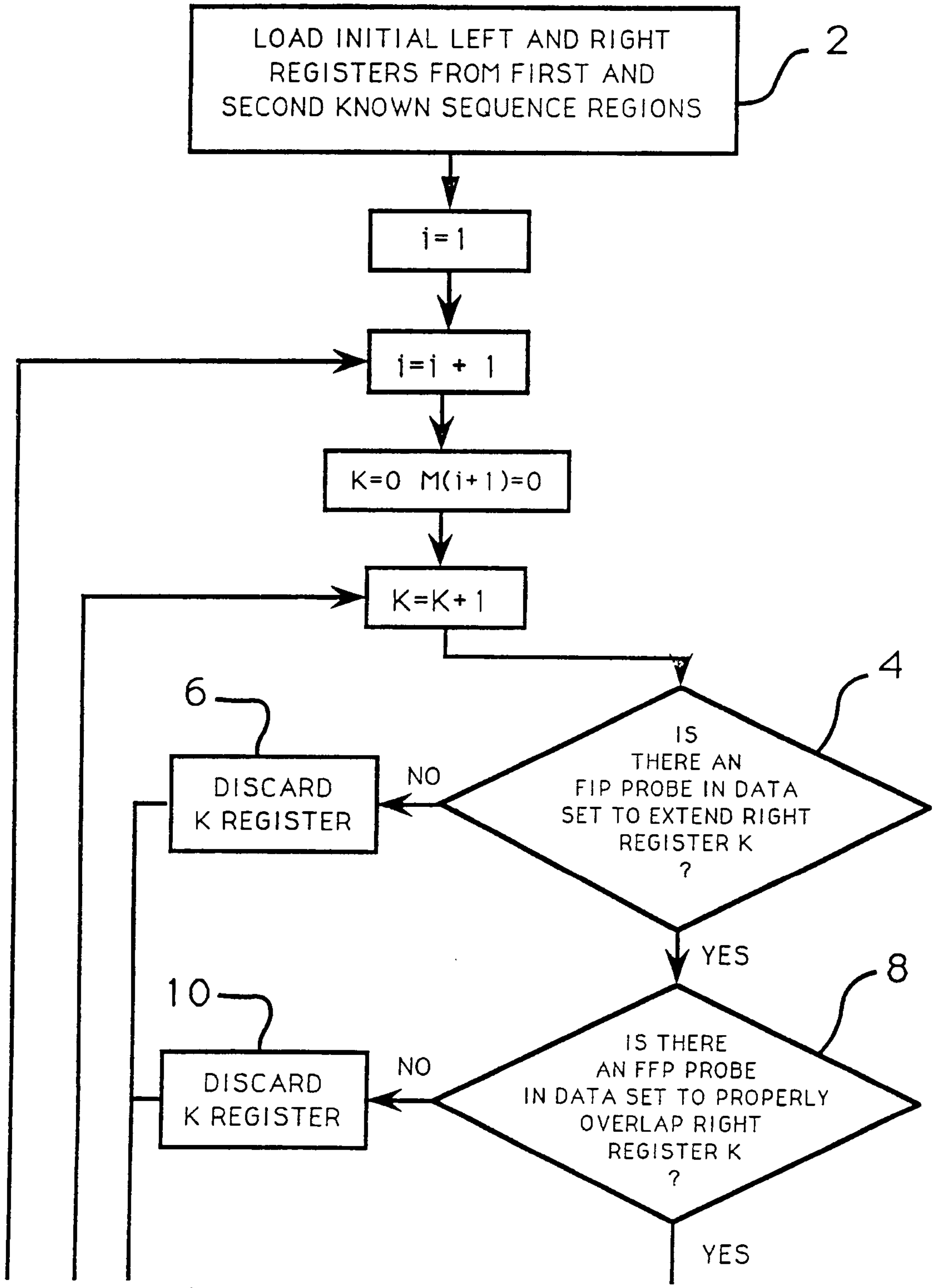


FIGURE 2 (part 1)

SUBSTITUTE SHEET

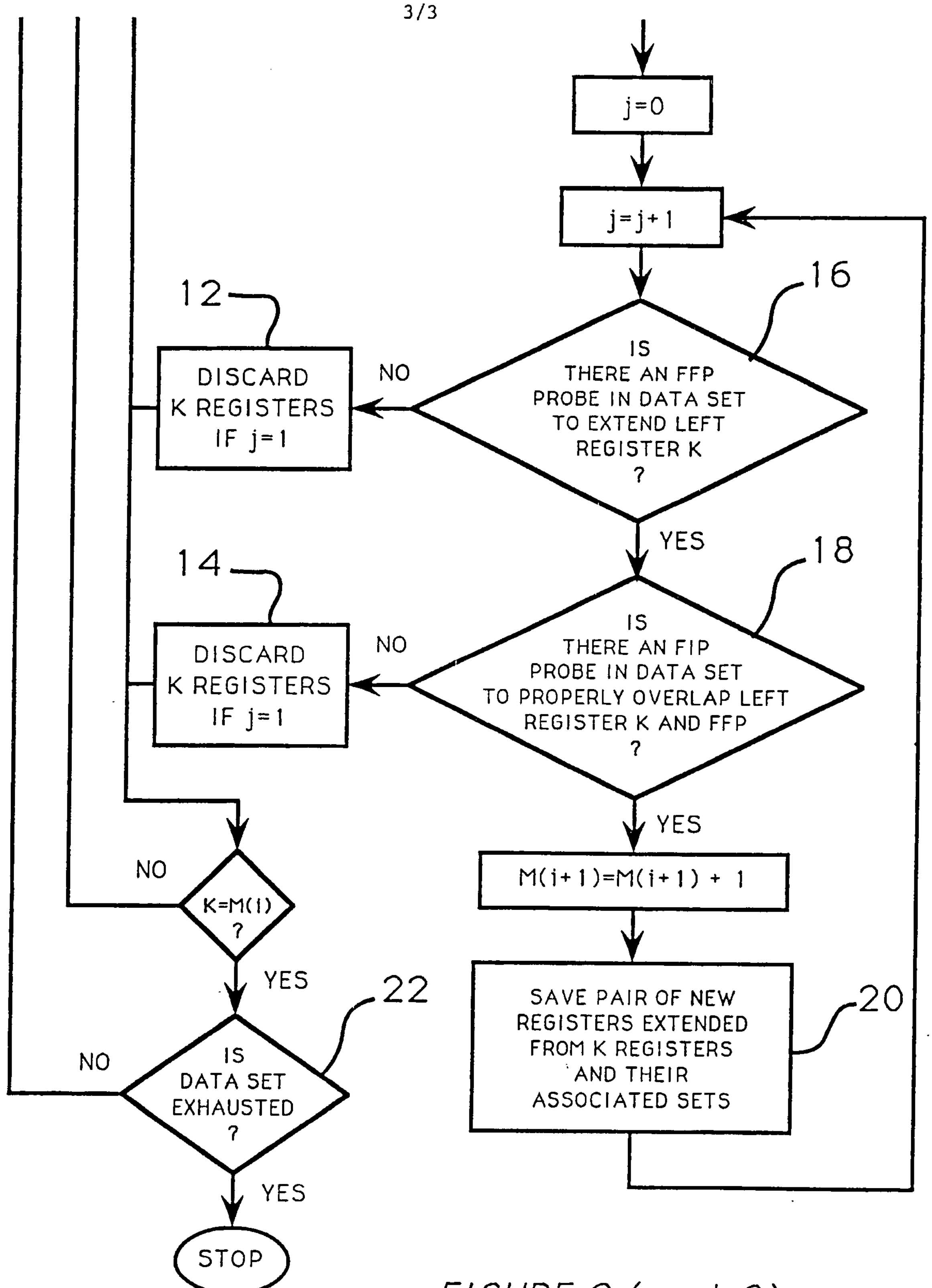


FIGURE 2 (part 2)

INTERNATIONAL SEARCH REPORT

International Application No. PCT/US89/04741

I. CLASSIFICATION OF SUBJECT MATTER (if several classification symbols apply, indicate all) [6]

According to International Patent Classification (IPC) or to both National Classification and IPC

IPC(5): C12Q 1/68
U.S.CL.: 435/6

II. FIELDS SEARCHED

Minimum Documentation Searched [7]

Classification System	Classification Symbols
U.S. CL.	435/6, 436/501, 935/77,78

Documentation Searched other than Minimum Documentation to the Extent that such Documents are Included in the Fields Searched [8]

DNAX RESEARCH INSTITUTE OF MOLECULAR AND CELLULAR BIOLOGY, INC.

III. DOCUMENTS CONSIDERED TO BE RELEVANT [9]

Category *	Citation of Document, [11] with indication, where appropriate, of the relevant passages [12]	Relevant to Claim No. [13]
Y	US, A, 4,613,566 POTTER, 23 September 1986, (See claims)	1-11
Y,P	US, A, 4,868,104, KURN et al., 19 September 1989, (See abstract).	1-11
Y	US, A, 4,710,465, WEISSMAN et al., 1 December 1987, (See claims)	1-11
Y	US, A, 4,689,295, TABER et al., 25 August 1987, (See claims and examples).	1-11
Y,P	US, A, 4,849,332, LORINCZ, 18 July 1989, (See claims).	1-11

* Special categories of cited documents: [10]

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

IV. CERTIFICATION

Date of the Actual Completion of the International Search	Date of Mailing of this International Search Report
06 FEB 1990	05 MAR 1990
International Searching Authority	Signature of Authorized Officer
ISA/US	Amelia B. Yarbrough AMELIA BURGESS YARBROUGH